DOCUMENT RESUME

ED 291 757 TM 011 042

AUTHOR Kirby, Peggy C.; Oescher, Jeffrey
TITLE Testing for Critical Thinking: Improving Test Development and Evaluation Skills of Classroom Teachers.
PUB DATE Nov 87
NOTE 19p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Mobile, AL, November 11-13, 1987).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Cognitive Measurement; *Critical Thinking; Evaluative Thinking; High Schools; *Item Analysis; Material Development; *Mathematics Tests; *Science Tests; Secondary School Mathematics; Secondary School Science; Teacher Characteristics; *Teacher Made Tests; *Test Construction

ABSTRACT

A study was made to determine characteristics of teacher-composed classroom tests, with emphasis placed on describing the levels of knowledge addressed by the test items. In this preliminary investigation, 19 mathematics and 16 science teachers working in 4 high schools in a mixed suburban/rural school district were asked to: (1) complete a brief instrument describing the format, objectives, analysis, and uses of their tests as well as their level of confidence in their testing skills; and (2) supply the researchers with their most recently administered unit or quarter examination. A rating form was devised to analyze a sample of teacher-composed tests. Interrater agreement for a sample of the tests ranged from 90 to 100 percent. Teachers' perceptions of the levels of knowledge addressed by their test items were compared to the researchers' actual ratings by means of t-tests or mean differences with the alpha levels adjusted using Bonferroni's formula. Multivariate analyses of variance were used to examine the main effects of school and subject taught on the percentage of items addressing each level of knowledge. Results provide insights into teachers' perceived purposes for testing, construction of test items, cognitive levels tested, overall test presentation, and confidence in testing skills. Major weaknesses discovered include a tendency to test at low cognitive levels, flaws in construction of individual test items, and inadequate instructions. Study data are presented in seven tables. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED291757

Testing for Critical Thinking:
Improving Test Development and Evaluation Skills of Classroom Teachers

Peggy C. Kirby
Jeffrey Oescher

Department of Educational Leadership and Foundations
University of New Orleans
New Orleans, LA 70148
(504) 286-6169

Paper presented at the annual meeting of
the Mid-South Educational Research Association
November 1987

TM 011 042

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Peggy Kirby
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Introduction

Schools are by definition designed to be places of learning, places where improved student achievement is the major objective. The measurement of achievement levels plays an important part in efforts to accomplish this objective. It provides the type of formal and informal feedback necessary to make informed instructional and evaluative decisions. It is imperative that the tests used to measure achievement be as technically sound as is possible.

The measurement community has made great strides in providing the technical background necessary to accurately and reliably measure achievement. Most of these efforts have focused on large-scale, standardized testing programs (Stiggins & Bridgeford, 1985). Unfortunately, little is known about the assessments made on a classroom level by individual teachers (Lazar-Morris, Polin, May, & Barry, 1980). Research has shown that teachers in elementary and secondary schools have little pre-service or in-service exposure to measurement concepts. Courses in tests and measurement in most states, including Louisiana, are not typically required for teacher certification. Yet, the classroom teacher is responsible for the development and analysis of the majority of the tests to which students are exposed.

Results presented here were obtained in a preliminary investigation of the characteristics of actual teacher-made tests. Particular emphasis was placed on describing the levels of knowledge addressed by teacher-composed test items. The results support earlier findings of Stiggins and Bridgeford (1985), Fleming and Chambers (1983), and Gullickson and Ellwein (1985). In addition, data collected on teachers' perceptions of their own testing practices corroborated these findings. Thus, two sources of data substantiate the immediate need to examine and improve teacher-composed classroom tests, particularly as related to the observed paucity of items targeted at higher cognitive levels. If cultivation of higher order thinking skills is a desired educational outcome, then the primary tool used in student assessment appears seriously flawed. This study attempted to validate earlier empirical studies of teacher-made tests and to suggest possible causes of consistently

identified weaknesses.

Review of the Literature

There are three dominant methods being used to assess student achievement: standardized tests (including curriculum based tests and questions), teacher-made objective tests, and observation-based assessments. The relative importance of these methods to the classroom teacher is clear - teacher-made tests and observational methods account for an overwhelming proportion of classroom assessments.

Stiggins and Bridgeford (1985) indicate that objective teacher-made tests are used most frequently, regardless of assessment purpose. Observed performance assessments, both structured and spontaneous, are the second most frequently used method. Published tests, including standardized objective achievement tests and objective tests supplied as part of textbook materials, play a secondary role. Salmon-Cox (1982) reported similar findings, with observational methods accounting for slightly more weight than teacher-made objective tests.

Research on the characteristics of teacher-made assesments has been scarce. Fleming and Chambers (1983) examined 342 tests that included over 8,800 items. They indicated that teacher-made tests tend to use short answer and matching item formats. Over 75% of all of the questions on the tests they examined were written in these formats. Multiple choice and true-false formats accounted for 14% and 10% respectively of all questions. Essays, even in English classes, accounted for less than 1% of the item formats.

According to Fleming and Chambers (1983), most test questions were written to examine low level cognitive skills (see Bloom, Madaus, & Hastings, 1981). Approximately 80% of all of the items focused on levels synonymous with the knowledge level of Bloom's Taxonomy. Comprehension level questions, including those examining skills in using processes and procedures and those requiring the ability to make translations, accounted for 17% of all items. Only 3% of all items were written at the application level. Detailed examination of the data indicated a tendency for junior high school teachers to use relatively more knowledge level questions than elementary and high school teachers. Also, math teachers tended to vary among behavioral levels more so than teachers of other subjects.

Fleming and Chambers' (1983) research identified several consistent problems with teacher-made tests. Of these, the most disconcerting is the lack of test items addressing higher order thinking skills. Only 3% of the items focused on skills at the application level. Although Fleming and Chambers' instruments did not include classifications for the analysis, evaluation, and synthesis levels, they concluded the "virtual absence of questions targeted at [the application level] suggests that instructional priorities are placed elsewhere" (p. 36). Other consistent but less troublesome problems that were identified included ambiguous short answer items, poor arrangement of multiple choice options, grammatical errors, lack of test directions, and failure to include point values for items.

Empirical evidence suggests that the lack of items targeted at higher order thinking skills is a function of teachers' inability to apply proven test development skills. Research conducted by Carter (1984) examining teachers' understanding of measurement principles found that only 30% of the responding teachers could correctly identify the level of items addressing higher order behaviors. When asked to write questions to address these levels, they required more time, had greater difficulty, and were less accurate than when writing lower order items.

Gullickson and Ellwein's (1985) research is consistent with the proposition that a measurement problem exists at the classroom level. Their survey of 150 elementary, junior high, and high school teachers indicated that few, if any, empirical analyses of test results are performed by classroom teachers. Although many teachers reported calculating reliability and difficulty indexes, an in-depth analysis of the researchers' instrument failed to support such claims. It was obvious that a wide gap existed between those skills prescribed by measurement specialists and those used by the classroom teacher.

Teacher-made objective tests are major sources of information used by classroom teachers. This is especially true for secondary math and science teachers. In general, these tests need improvement, particularly in the areas of the level of behavior addressed by the items and the analysis of test results. Teachers tend to feel somewhat insecure when working with these topics, indicating a need for programs that will offer practical advice on the use and

application of these and other measurement principles. Unfortunately, programs such as these exist only in isolated instances.

Methodology

Sample

All teachers of math and science at the senior high level (9th - 12th grades) in a mixed suburban/rural school district were asked to participate in a research project examining teacher testing practices. Thirty-five teachers - 19 math and 16 science teachers - from four high schools participated. Their involvement consisted of 1) responding to items of a brief instrument describing the format, objectives, analysis, and uses of their tests, as well as level of confidence in their testing skills; and 2) supplying the researchers with their most recently administered unit or quarter exam.

One teacher reported administering only oral exams. He, therefore, would participate only in the survey portion of the study. Subjects were guaranteed confidentiality in the reporting of results. Also, at the request of several participants, it was agreed that the sample tests would be returned subsequent to analysis.

Instrumentation

Teachers in the sample completed a brief instrument describing their testing practices. In addition to estimating the percentage of items written in each of five formats and at each of four levels of knowledge, subjects responded to items regarding their purposes and uses of classroom tests, their analyses of test results, and their perceived confidence in test development. This Teacher Testing Questionnaire was developed based on problems in measurement revealed through the review of the literature.

A rating form was devised to analyze a sample of teacher-composed tests. Raters recorded the number of items written in each of five formats; estimated the level of knowledge targeted by each item; judged the quality of multiple choice, true/false, matching, short answer, and essay items; and evaluated specific characteristics of the

overall presentation such as adequacy of instructions, numbering system, formatting, and duplication. Inter-rater agreement for a sample of these tests on all items where percent agreement was deemed an appropriate reliability indicator ranged from 90 to 100 percent.

Procedure

The district superintendent wrote a letter of support for the proposed research to each high school principal in his district. The researchers then met with the principals individually to request copies of each math and science teacher's most recently administered quarter exam. Teachers who did not administer quarter exams supplied their most recent unit test. Once tests were collected, teachers were asked to respond to items of the Teacher Testing Questionnaire and return it in a sealed envelope to the principal or a designee for forwarding to the research team. Thirty-four tests and 35 Teacher Testing Questionnaires were collected. Although some teachers supplied multiple tests, only one per teacher was chosen at random for analysis.

Data Analysis

The sample of tests was scored by the researchers assigning percentages for levels of knowledge and item formats, and one to three ratings to items describing overall presentation and quality of items in each format.

Descriptive statistics were computed for items of both the Teacher Testing Questionnaire and the actual test analyses using SAS Release 5.16 (1985), a statistical software package. Teachers' perceptions of the levels of knowledge addressed by their test items were compared to the researchers' actual ratings by means of *t*-tests of mean differences with the alpha levels adjusted using Bonferroni's formula (Dunn, 1961). Multivariate analyses of variance (MANOVA's) were used to examine the main effects of school and subject taught on the percentage of items addressing each level of knowledge.

Results

Purposes for Testing

Thirty-one of 35 (94%) of the teachers surveyed reported that they place the most emphasis in student evaluation on their own classroom tests. Classroom participation and effort ranked second in relative importance while standardized tests carried the least weight. The other choices, feedback obtained from instruction and student behavior, are given relatively minor emphasis in assigning student grades. These findings confirm Stiggins and Bridgeford's conclusion that teacher-made tests account for a large proportion of classroom assessments.

The most commonly reported purpose for testing was assignment of student grades. Teachers said that 71.9% of all tests were administered for this purpose. On average, only 12% of their tests were used to evaluate instruction, and less than 2% were used for placement (see Table 1). Teachers, however, claimed to frequently review tests with their students, identify student weaknesses and modify instruction based on test results (see Table 2). Thus, there appears to be recognition of various uses of test results but emphasis on the summative rather than formative role in student assessment.

Construction of Test Items

Item Format. Teachers were asked to estimate the percent of items they write in each of five formats: multiple choice, true/false, matching, short answer (including fill-in-the-blank), and essay. Additionally, raters sorted items from the 34 sample tests to these categories. All fill-in items with a supplied list of choices were considered as matching items.

Teachers' perceptions are compared to observed percentages in Table 3. Teachers reported that the most often used format was short answer with over 40% of all items written in this format. Our analyses revealed that over 60% of all items were actually of the short answer variety. Teachers said that matching and multiple choice formats accounted for 15.5% each of all items. Of the test items analyzed, 19.0% were multiple choice and 15.6% were matching. True/false comprised 8.3% of all items; teachers

estimated that this format was used in approximately 5.0% of all items. Although teachers reported that one in every five items was written in essay format, our analyses revealed only four essay items in over 1400 items examined.

Teachers do not routinely weight item formats differently. As revealed in Table 4, the percentage of a student's score determined by any one item format parallels the percentage of items written in that format.

Flaws were detected in the majority of teacher-composed test sections. Raters judged groups of items in similar format for each test as containing errors in more than 20% of the items, in 20% or less items, or in no items. Of the 18 tests containing multiple choice items, 17 were judged to have flaws in more than 20% of these items. More than 20% of the true/false items on five of ten tests were determined to be poorly written. Matching items were weak on 11 of 12 tests, and short answer items were judged poor on 21 of 29 tests containing this format. Of the four essay items presented, all contained major flaws. None of these contained information to guide the student in structuring a response or tapped higher level thinking skills.

Cognitive Levels Tested. Teachers agree that the vast majority of items are written at the lower cognitive levels of knowledge and comprehension (Bloom, Madaus, & Hastings, 1981). A major discrepancy lies, however, in the perceived percentage of items written at higher levels. Although teachers report that roughly one-fourth of all items are written at the application, analysis, synthesis, or evaluation level, our analyses place less than 8% of all items at these cognitive levels, with virtually no items requiring students to synthesize or evaluate. A t-test of mean differences between teacher perceptions of the percentage of items written at the levels of synthesis or evaluation and rater judgments of percentage of items at these levels was statistically significant ($t=4.76$, $p<.001$ with Bonferroni correction). This discrepancy confirms Carter's (1984) finding that teachers tend to inaccurately classify higher order items.

Possible effects of school and subject taught were analyzed. Results indicated that the individual school had no effect on teachers' use of higher level test items. However, the subject - math or science - did significantly effect the percentage of items judged to be written at the knowledge and comprehension levels. No differences by

subject were found at other cognitive levels. Although teachers of both disciplines write the majority of items at these two lower levels, math teachers include significantly greater numbers of comprehension items on their exams (see Table 6). While science tests analyzed contained, on average, 78.2% of all items at the knowledge level and 16.8% at the comprehension level, math tests had an average of 77.5% of all items written at the comprehension level with 12.5% at the knowledge level. This finding can be attributed to the tendency to test math skills by requiring students to use rules or procedures (comprehension level) to solve number problems.

The finding of major importance here is not the differences by subject at the lower levels of knowledge, but the lack of items in either subject at higher levels. Interestingly, few math teachers required students to apply knowledge of procedures to new situations. Word problems were regretfully scarce.

Overall Test Presentation

Teachers, on average, reported writing 65.6% of their test items themselves with the remaining items being obtained from test guides, textbooks, workbooks, and other sources. Grammatical errors discovered by raters were relatively few in number. Three of the tests analyzed (8.8%) were judged to contain many grammatic errors, six (17.6%) contained few errors, and 25 (73.5%) contained no errors.

The average number of items per test was 42.0 with a standard deviation of 23.7. Number of items varied widely with a minimum of 14 items and a maximum of 103.

Twenty-four of the tests were completely type-written, two contained both typed and hand-written sections, and eight were totally hand-written. In only four cases was duplication quality judged to be inadequate. These tests were deemed to be "readable but with difficulty." Formatting was a problem in over 70% of the tests analyzed. These deficiencies consisted of crowding, inconsistent style or margins, and lack of space for answers.

Students were not typically informed of the point value of any test or test item. None of the 34 tests analyzed

contained a written explanation of the weight of that test in determining a student's grade. In only six tests (17.6%) was the point value of individual items or sections provided. On average, however, teachers reported that they frequently informed students of item values. Yet empirical evidence suggests that students were not aware of the relative emphasis attributed to any item or section unless this information was verbalized to them prior to testing.

Written instructions were provided on 25 of the 34 tests. All but two of these contained instructions for the total test as well as subsections. Nine tests (26.5%) contained no instructions despite the fact that teachers reported nearly always including instructions for each subsection (see Table 7).

Instructions were deemed "nebulous" for 21 of the 25 tests (84.0%) that contained written instructions. "Nebulous" was used to refer to instructions such as those that ask students to choose an answer without indicating how or where the choice is to be recorded. This was particularly problematic for matching items where two long lists were often presented with no space provided for answers. The student was left to decide whether to match Column A to Column B, Column B to Column A, or draw lines between the two.

Teachers' Confidence in Testing Skills

Teachers were asked to respond on a scale of 1 to 5 with 1 equal to "strongly disagree" and 5 equal to "strongly agree" how confident they felt in their testing skills. They reported, on average, feeling confident in their ability to construct valid and reliable tests ($\bar{M}$=4.40) and assess the validity and reliability of those tests ($\bar{M}$=4.29). They tended to rate their pre-service training in tests and measurement as adequate ($\bar{M}$=3.71) and were only slightly less assured of the adequacy of their in-service training ($\bar{M}$=3.49).

In spite of these perceptions, many commonly accepted test development and analysis procedures were not routinely practiced (see Tables 7 and 8). Teachers did report frequently using an answer key in scoring objective test items and writing out desired responses before scoring essay items. They tended to determine point values for items before correcting tests. However, they did not eliminate

poor items based on test results. They reported only occasionally computing item analysis information or even an arithmetic mean of test scores.

Teachers claimed to very frequently base tests on instructional objectives, but only occasionally tallied the number of items per objective or per skill level. Thus, tables of specification do not appear to be used on a regular basis.

Conclusions Regarding Teacher Skill in Test Development

The major weaknesses noted from analyses of these same teachers' tests were tendency to test at low cognitive levels, flaws in the construction of individual test items, and inadequate instructions. Testing is a major area of concern for parents, students, and teachers. Scores determined from the results of teacher-made tests directly affect student grades and placement, yet the reliability of classroom tests is questionable given the observed flaws in item writing and presentation of instructions, as well as the failure of most teachers to calculate item analysis information. The content validity of these tests is also uncertain since only a small range of knowledge is addressed by their items, and because a table of specifications, a major tool of valid test development, is not commonly used.

Research describing the characteristics of teacher-made tests has uncovered recurring problems. It is imperative that we now begin to train teachers in principles of test construction, particularly as related to higher order thinking skills. The current public emphasis on tests and measurements demands that this aspect of testing, the aspect that accounts for the largest proportion of student assessments, be improved.

REFERENCES

Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). Evaluation to improve learning. New York: McGraw-Hill.

Carter, K. (1984). Do teachers understand principles for writing tests? Journal of Teacher Evaluation. 35, 57-60.

Dunn, O.J. (1961). Multiple comparisons among means. Journal of the American Statistical Association, 56, 52-64.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W.E Hathaway (Ed.), Testing in the Schools: New Directions for Testing and Measurement, 19, (pp. 29-38). San Francisco: Jossey-Bass.

Gullickson, A.R., & Ellwein, M.C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, 4, 15-18.

Lazar-Morris, C., Polin, L., May, R., & Barry, L. (1980). A review of the literature on test use. Los Angeles: University of California, Center for the Study of Evaluation. (ERIC Document Reproduction Service No. ED 204 411)

Salmon-Cox, L. (1982). Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 62, 631-634.

SAS Institute, Inc. (1985). SAS [Computer program]. Carey, NC: Author.

Stiggins, R.J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22, 271-286.

Table 1

Teachers' Reported Purposes for Testing

% of tests used for diagnosis	% used for placement of students	% used for assigning grades	% used for evaluating instruction	% used for reinforcing instructio
5.97	1.42	71.90	12.03	10.13

Table 2

How Teachers' Report Using Test Results

	Review tests with students	Identify student weaknesses	Modify instruction	Assign remedial or supplemental work
M*	4.57	4.26	3.86	3.00
SD	.74	.66	.65	.77

Note. 1 = Never
2 = Seldom
3 = Sometimes
4 = Frequently
5 = Always

Table 3

Item Formats

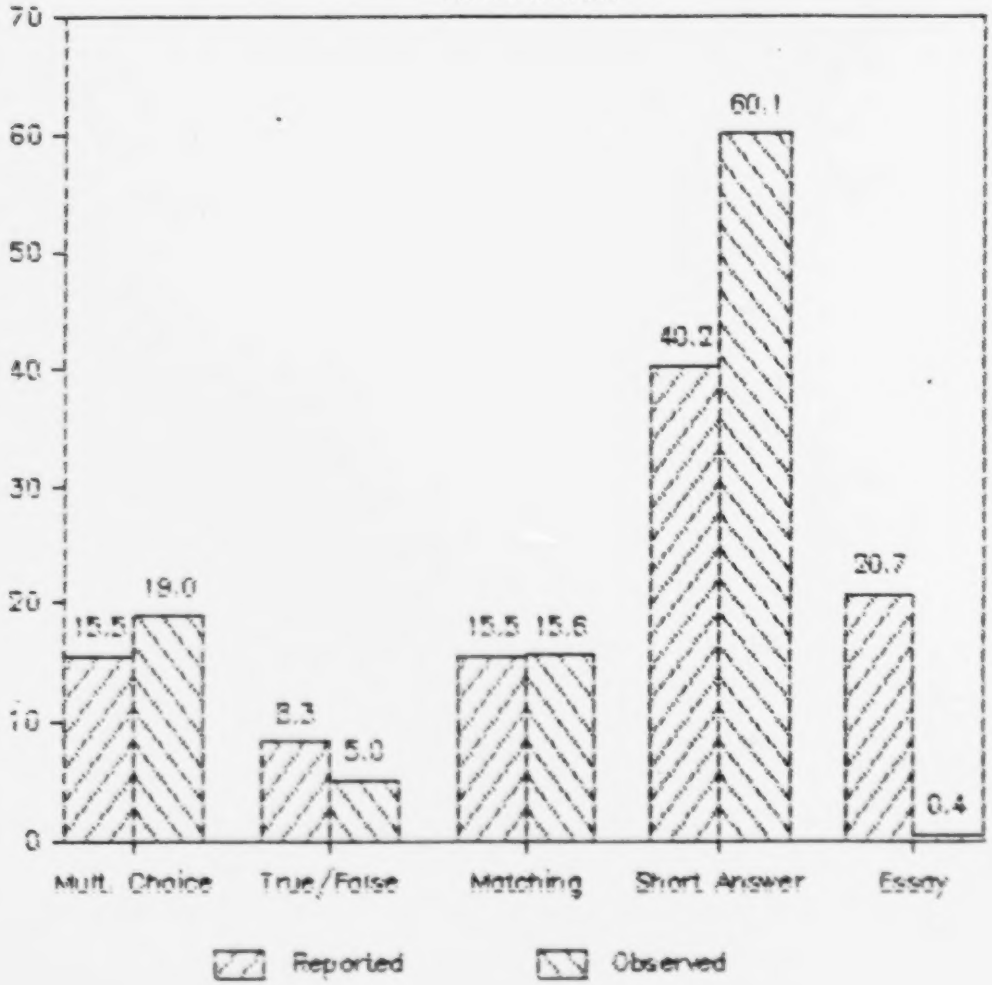

Table 4

Percent of Student Score Obtained From Each Item Format

	Multiple Choice	True/ False	Matching	Short Answer	Essay
Teacher-reported % of items in each format	15.46	8.31	15.46	40.23	20.69
Teacher-reported % of score obtained from each format	13.68	8.12	14.12	41.76	22.32

Table 5

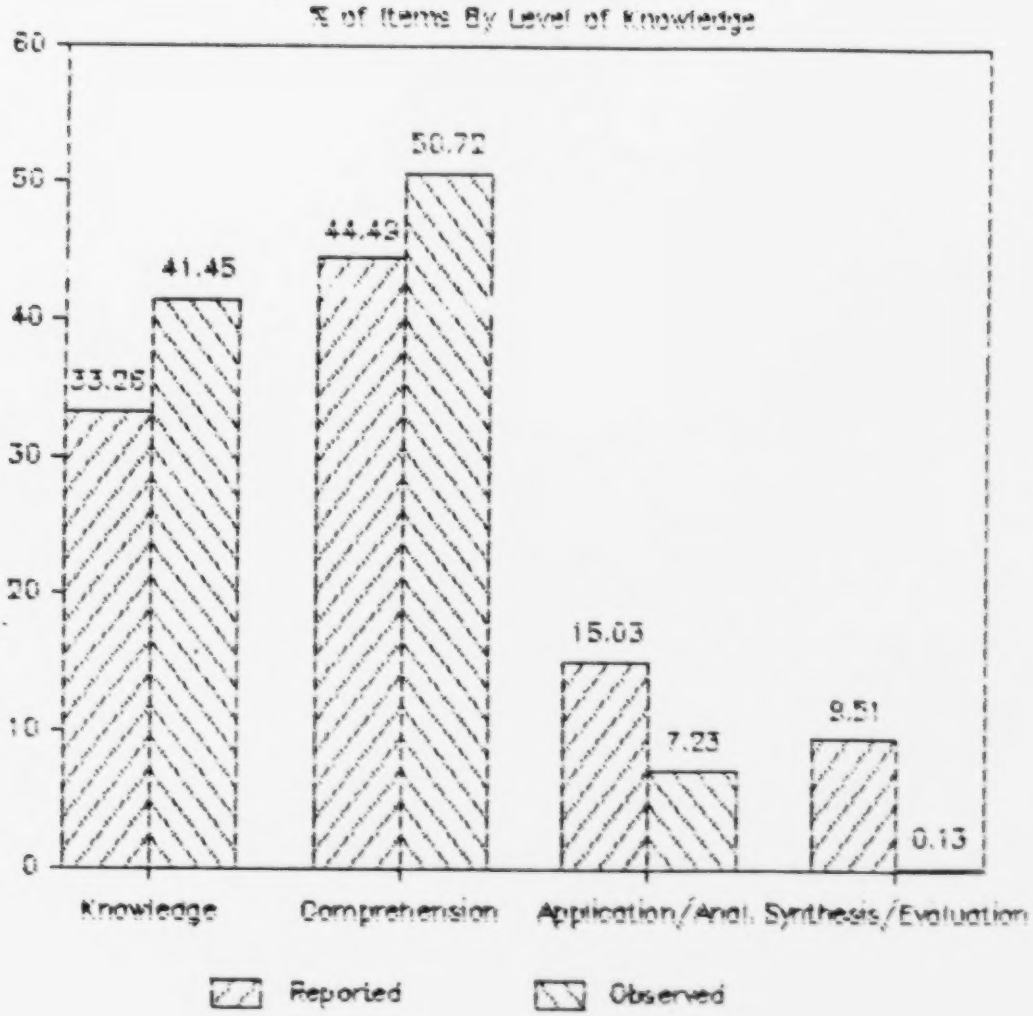
% of Items By Level of Knowledge
60
50
40
30
20
10
0
33.26
41.45
44.43
50.72
15.03
7.23
9.51
0.13
Knowledge
Comprehension
Application/Anal. Synthesis/Evaluation
Reported
Observed

Table 6

ANOVA Summary Tables for Effect of Subject Taught on % of Items Observed at Knowledge and Comprehension Levels

Dependent Variable: Knowledge

	df	SS	F
Subject	1	3.49	94.04*
Error	32	1.39	
Total	33	4.88	

* $p < .0001$

Dependent Variable: Comprehension

	df	SS	F
Subject	1	2.84	74.63*
Error	32	1.57	
Total	33	4.41	

* $p < .0001$

Table 7

Reported Testing Practices of 35 Math and Science Teachers

Item	M	SD
My tests are based on my instructional objectives	4.85	.36
I tally the number of items intended to measure each instructional objective	3.31	1.18
I tally the number of items intended to measure each level of student performance	2.97	1.15
I include written instructions for each section of my tests	4.60	.91
My students are informed of the point value of each test item	4.06	1.00
I complete an answer key for each objective item before scoring tests	4.80	.63
I write out an appropriate or desired response for each essay item before scoring these items	4.38	1.15
Scores on my tests are adjusted for guessing	1.76	1.23
I assign the point values for individual items before correcting all tests	3.21	1.01
I compute item analysis information for my tests	2.36	.90
I eliminate certain items in determining test scores	2.42	.61
I compute an arithmetic mean of scores received by students for each test	2.63	1.21

Note. 1 = Never 2 = Seldom 3 = Sometimes
4 = Frequently 5 = Always